Confirmation and Induction

Confirmation and Induction

Jan Sprenger*

Abstract

Scientific knowledge is based on induction, that is, ampliative inferences from experience. This chapter reviews attacks on and defenses of induction, as well as attempts to spell out rules of inductive inference. Particular attention is paid to the *degree of confirmation* of a hypothesis, its role in inductive inference and the Bayesian explications of that concept. Finally, the chapter compares Bayesian and frequentist approaches to statistical inference.

Keywords: confirmation, degree of confirmation, induction, probability, inductive logic, Bayesianism, statistical inference.

*Contact information: Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

1 The Problems of Induction

Induction is a method of inference that aims at gaining empirical knowledge. It has two main characteristics: First, it is **based on experience**. (The term "experience" is used interchangeably with "observations" and "evidence".) Second, induction is **ampliative**, that is, the conclusions of an inductive inference are not necessary, but contingent. The first feature makes sure that induction targets empirical knowledge, the second feature distinguishes induction from other modes of inference, such as deduction, where the truth of the premises guarantees the truth of the conclusion.

Induction can have many forms. The most simple one is **enumerative induction**: inferring a general principle or making a prediction based on the observation of particular instances. For example, if we have observed 100 black ravens and no non-black ravens, we may predict that also raven #101 will be black. We may also infer the general principle that all ravens are black. But induction is not tied to the enumerative form and comprises all ampliative inferences from experience. For example, making weather forecasts or predicting economic growth rates are highly complex inductive inferences that amalgamate diverse bodies of evidence.

The first proper canon for inductive reasoning in science has been set up by Francis Bacon, in his *Novum Organon* (Bacon, 1620). Bacon's emphasis is on learning the cause of a scientifically interesting phenomenon. He proposes a method of **eliminative induction**, that is, eliminating potential causes by coming up with cases where the cause, but not the effect is present. For example, if the common flu occurs in a hot summer period, then cold cannot be its (sole) cause. A similar method, though with less meticulous devotion to the details, has been outlined by René Descartes (1637). In his *Discours de la Méthode*, he explains how scientific problems should be divided into tractable subproblems, and how their solutions should be combined.

Both philosopers realize that without induction, science would be blind to experience and unable to make progress. Hence their interest in spelling out the inductive method in detail. However, they do not provide a foundational *justification* of inductive inference. For this reason, C.D. Broad (1952, 142–143) stated that "inductive reasoning [...] has long been the glory of science", but a "scandal of philosophy". This quote brings us directly to the notorious **problem of induction** (for a survey, see Vickers, 2010).

Two problems of induction should be distinguished. The first, fundamental problem is about why we are justified to make inductive inferences, that is, why the method of induction works at all. The second problem is about telling good from bad inductions and developing **rules of inductive inference**. How do we learn from experience? Which inferences about future predictions or general theories are justified by these ob-

servations? And so on.

About 150 years after Bacon, David Hume (1739, 1748) was the first philosopher to clearly point out how hard the first problem of induction is (*Treatise On Human Nature*, 1739, Book I; *Enquiry Concerning Human Understanding*, 1748, Sections IV+V). Like Bacon, Hume is interested in learning the causes of an event as a primary means of acquiring scientific knowledge. Since causal relations cannot be inferred *a priori*, we have to learn them from experience, that is, to use induction.

Hume divides all human reasoning into demonstrative and probabilistic reasoning. He notes that learning from experience falls into the latter category: no amount of observations can logically *guarantee* that the sun will rise tomorrow, that lightning is followed by thunder, that England will continue to lose penalty shootouts, etc. In fact, regularities of the latter sort sometimes cease to be true. Inductive inferences cannot *demonstrate* the truth of the conclusion, but only make it *probable*.

This implies that inductive inferences have to be justified by non-demonstrative principles. Imagine that we examine the effect of heat on liquids. We observe in a number of experiments that water expands when heated. We predict that upon repetition of the experiment, the same effect will occur. However, this is probable only if nature does not change its laws suddenly: "all inferences from experience suppose, as their foundation, that the future will resemble the past" (Hume, 1748, 32). We are caught in a vicious circle: the justification of our inductive inferences invokes the principle of induction itself. This undermines the rationality of our preference for induction over other modes of inference, e.g., counter-induction.

The problem is that assuming the uniformity of nature in time can only be justified by inductive reasoning, namely our past observations to that effect. Notably, also pragmatic justifications of induction, by reference to past successes, do not fly since inferring from past to future reliability of induction also obeys the scheme of an inductive inference (ibid.).

Hume therefore draws the skeptical conclusion that we lack a rational basis for believing that causal relations inferred from experience are necessary or even probable. Instead, what makes us associate causes and effects are the irresistisible psychological forces of custom and habit. The connection between cause and effect is in the mind rather than in the world, as witnessed by our inability to give an independent rational justification of induction (Hume, 1748, 35–38).

Hume's skeptical argument seems to undermine a lot of accepted scientific method. If induction does not have a rational basis, why perform experiments, predict future events and infer to general theories? Why science at all? Note that Hume's challenge also affects the second problem: if inductive inferences cannot be justified in an objective way, how are we going to tell which rules of induction are good and which are

bad?

Influenced by Hume, Karl Popper (1959, 1983) developed a radical response to the problem of induction. For him, scientific reasoning is essentially a deductive and not an inductive exercise. A proper account of scientific method neither affords nor requires inductive inference—it is about **testing hypotheses** on the basis of their predictions:

> The best we can say of a hypothesis is that up to now it has been able to show its worth [...] although, in principle, it can never be justified, verified, or even shown to be probable. This appraisal of the hypothesis relies solely upon deductive consequences (predictions) which may be drawn from the hypothesis: There is no need even to mention "induction". (Popper, 1959, 346)

For Popper, the merits of a hypothesis are not determined by the degree to which past observations support it, but by its performances in severe tests, that is, sincere attempts to overthrow it. Famous examples from science include the Michelson-Morley experiment as a test of the ether theory and the Allais and Ellsberg experiments as tests of Expected Utility Theory. Popper's account also fits well with some aspects of statistical reasoning, such as the common use of **Null Hypothesis Significance Tests (NHST)**: a hypothesis of interest is tested against a body of observations and "rejected" if the result is particularly unexpected. Such experiments do not warrant inferring or accepting a hypothesis; they are exclusively designed to *disprove* the null hypothesis and to collect evidence against it. More on NHST will be said in Section 6.

According to Popper's view of scientific method, induction in the narrow sense of inferring a theory from a body of data is not only unjustified, but even superfluous. Science, our best source of knowledge, assesses theories on the basis of whether their predictions obtain. Those predictions are deduced from the theory. Hypotheses are *corroborated* when they survive a genuine refutation attempt, when their predictions were correct. Degrees of corroboration may be used to form practical preferences over hypotheses. Of course, this also amounts to learning from experience and to a form of induction, broadly conceived—but Popper clearly speaks out against the view that scientific hypotheses with universal scope are ever guaranteed or made probable by observations.

Popper's stance proved to be influential in statistical methodology. In recent years, philosopher Deborah Mayo and econometrist Aris Spanos have worked intensively on this topic (e.g., Mayo, 1996; Mayo and Spanos, 2006). Their main idea is that our preferences among hypotheses are based on the *degree of severity* with which they have been tested. Informally stated, they propose that a hypothesis has been severely

tested if (i) it fits well with the data, for some appropriate notion of fit; and (ii) if the hypothesis were false, it would have been very likely to obtain data that favor the relevant alternative(s) much more than the actual data do.

We shall, however, not go into the details of their approach and return to the second problem of induction: how should we tell good from bad inductions?

2 Logical Rules for Inductive Inference

Hume's skeptical arguments show how difficult it is to argue for the reliability and truth-conduciveness of inductive inference. However, this conclusion sounds more devastating than it really is. For example, on a reliabilist view of justification (Goldman, 1986), beliefs are justified if generated by reliable processes that usually lead to true conclusions. If induction is factually reliable, our inductive inferences are justified even if we cannot access the reasons for why the method works. In a similar vein, John Norton (2003) has discarded formal theories of induction (e.g., those based on the enumerative scheme) and endorsed a *material* theory of induction: inductive inferences are justified by their conformity to facts.

Let us now return to the second problem of induction, that is, developing (possibly domain-sensitive) **rules of induction**—principles that tell good from bad inductive inferences. In developing these principles, we will make use of the method of **reflective equilibrium** (Goodman, 1955): we balance scientific practice with normative considerations, e.g., which methods track truth in the idealized circumstances of formal models. Good rules of induction are those that explain the success of science and that have at the same time favorable theoretical properties. The entire project is motivated by the analogy to deductive logic, where rules of inference have been useful at guiding our logical reasoning. So why not generalize the project to **inductive logic**, to rules of reasoning under uncertainty and ampliative inferences from experience?

Inductive logic has been the project of a lot of 20th century philosophy of science. Sometimes it also figures under the heading of finding criteria for when evidence confirms (or supports) a scientific hypothesis. The presence of a confirmation relation, or the degree to which a hypothesis is confirmed, provides a criterion for the soundness of an inductive inference. It is therefore sensible to **explicate the concept of confirmation**: to replace our vague pre-theoretical concept, the *explicandum*, with a simple, exact and fruitful concept that still resembles the explicandum—the *explicatum* (Carnap, 1950, 3–7). The explication can proceed **quantitatively**, specifying degrees of confirmation, or **qualitatively**, as an all-or-nothing relation between hypothesis and evidence. We will first look at qualitative analyses in first-order logic since they outline

the logical grammar of the concept. Several features and problems of qualitative accounts carry over to and motivate peculiar quantitative explications (Hempel, 1945a).

Scientific laws often take the logical form $\forall x : Fx \rightarrow Gx$, that is, all F's are also G's. For instance, take Kepler's First Law that all planets travel in an elliptical orbit around the sun. Then, it is logical to distinguish two kinds of confirmation of such laws, as proposed by Jean Nicod (1961, 23–25): *L'induction par l'infirmation* proceeds by refuting and eliminating other candidate hypotheses (e.g., the hypothesis that planets revolve around the Earth). This is basically the method of eliminative induction that Bacon applied to causal inference. *L'induction par la confirmation*, by contrast, supports a hypothesis by citing their *instances* (e.g., a planet which has an elliptical orbit around the sun). This is perhaps the simplest and most natural account of scientific theory confirmation. It can be expressed as follows:

> **Nicod Condition (NC):** For a hypothesis of the form $H = \forall x : Fx \rightarrow Gx$ and an individual constant a, an observation report of the form $Fa.Ga$ confirms H.

However, (NC) fails to capture some essentials of scientific confirmation—see Sprenger (2010) for details. Consider the following highly plausible adequacy condition, due to Carl G. Hempel (1945a,b):

> **Equivalence Condition (EC):** If H and H' are logically equivalent sentences, then E confirms H if and only if E confirms H'.

(EC) should be satisfied by any logic of confirmation because otherwise, the establishment of a confirmation relation would depend on the peculiar formulation of the hypothesis, which would contradict our goal of finding a *logic* of inductive inference.

Combining (EC) with (NC) leads, however, to paradoxical results. Let $H = \forall x : Rx \rightarrow Bx$ stand for the hypothesis that all ravens are black. H is equivalent to the hypothesis $H' = \forall x : \neg Bx \rightarrow \neg Rx$ that no non-black object is a raven. A white shoe is an instance of this hypothesis H'. By (NC), observing a white shoe confirms H', and by (EC), it also confirms H. Hence, observing a white shoe confirms the hypothesis that all ravens are black! But a white shoe appears to be an utterly irrelevant piece of evidence for assessing the hypothesis that all *ravens* are black. This result is often called the **paradox of the ravens** (Hempel, 1945a, 13–15) or, after its inventor, **Hempel's paradox**.

How should we deal with this problem? Hempel suggests to bite the bullet and to accept that the observation of a white shoe confirms the raven hypothesis. After all, the observation eliminates a potential falsifier. To push this intuition further, imagine that we observe a grey, raven-like bird, and only after extended scrutiny we find out

that it is a crow. There is certainly a sense in which the crowness of that bird confirms the raven hypothesis, which was already close to refutation.

Hempel (1945a,b) implements this strategy by developing a more sophisticated version of Nicod's instance confirmation criterion where background knowledge plays a distinct role, the so-called **Satisfaction Criterion**. We begin with the formulation of *direct* confirmation, which also captures the main idea of Hempel's proposal:

Direct Confirmation (Hempel) A piece of evidence E *directly Hempel-confirms* a hypothesis H *relative to background knowledge* K if and only if E and K jointly entail the development of H to the domain of E—that is, the restriction of H to the set of individual constants that figure in E. In other words, $E.K \models H_{|dom(E)}$.

The idea of this criterion is that our observation verify a general hypothesis, as restricted to the actually observed objects. Hempel's **Satisfaction Criterion** generalizes this intution by demanding that a hypothesis be confirmed whenever it is entailed by a set of directly confirmed sentences. Notably, Clark Glymour's account of *bootstrap confirmation* is also based on Hempel's Satisfaction Criterion (Glymour, 1980b).

However, Hempel did not notice that the Satisfaction Criterion does not resolve the raven paradox: $E = \neg Ba$ directly confirms the raven hypothesis H relative to $K = \neg Ra$ (because $E.K \models H_{\{a\}}$). Thus, even objects *that are known not to be ravens* can confirm the hypothesis that all ravens are black. This is clearly an unacceptable conclusion and invalidates the Satisfaction Criterion as an acceptable account of qualitative confirmation, whatever its other merits may be (Fitelson and Hawthorne, 2011).

Hempel also developed several **adequacy criteria** for confirmation, intended to narrow down the set of admissible explications. We have already encountered one of them, the Equivalence Condition. Another one, the Special Consequence Condition, claims that consequences of a confirmed hypothesis are confirmed as well. Hypotheses confirmed by a particular piece of evidence form a deductively closed set of sentences. The Satisfaction criterion conforms to this condition, as one can easily check from the definition. It also satisfies the Consistency Condition which demands (inter alia) that no contingent evidence supports two hypotheses which are inconsistent with each other. This sounds very plausible, but as noted by Nelson Goodman (1955) in his book "Fact, Fiction and Forecast", that condition conflicts with powerful inductive intuitions. Consider the following inference:

Observation: emerald e_1 is green.
Observation: emerald e_2 is green.
...

Generalization: All emeralds are green.

This seems to be a perfect example of a valid inductive inference. Now define the predicate "grue", which applies to all green objects if they were observed for the first time prior to time t = "now", and to all blue objects if observed later. (This is just a description of the extension of the predicate—no object is supposed to change color.) The following inductive inference satisfies the same logical scheme as the previous one:

Observation: emerald e_1 is grue.
Observation: emerald e_2 is grue.
...

Generalization: All emeralds are grue.

In spite of the gerrymandered nature of the "grue" predicate, the inference is sound: it satisfies the basic scheme of enumerative induction, and the premises are undoubtedly true. But then, it is paradoxical that two valid inductive inferences support flatly opposite conclusions. The first generalization predicts emeralds observed in the future to be green, the second generalization predicts them to be blue. How do we escape from this dilemma?

Goodman considers the option that in virtue of its gerrymandered nature, the predicate "grue" should not enter inductive inferences. He notes, however, that it is perfectly possible to re-define the standard predicates "green" and "blue" in terms of "grue" and its conjugate predicate "bleen" (=blue if observed prior to t, else green). Hence, any preference for the "natural" predicates and the "natural" inductive inference seems to be arbitrary. Unless we want to give up on the scheme of enumerative induction, we are forced into dropping Hempel's Consistency Condition, and to accept the paradoxical conclusion that both conclusions (all emeralds are green/grue) are, at least to a certain extent, confirmed by past observations. The general moral is that conclusions of an inductive inference need not be consistent with each other, unlike in deductive logic.

Goodman's example, often called the **new riddle of induction**, illustrates that establishing rules of induction and adequacy criteria for confirmation is not a simple business. From a normative point of view, the Consistency Condition looks appealing, yet, it clashes with intuitions about paradigmatic cases of enumerative inductive inference. The rest of this chapter will therefore focus on accounts of confirmation where inconsistent hypotheses can be confirmed simultaneously by the same piece of evidence.

A prominent representative of these accounts is **Hypothetico-Deductive (H-D) confirmation**. H-D confirmation considers a hypothesis to be confirmed if empirical predictions deduced from that hypothesis turn out to be successful (Gemes, 1998; Sprenger, 2011). An early description of H-D confirmation was given by William Whewell:

> Our hypotheses ought to *foretel* phenomena which have not yet been observed ... the truth and accuracy of these predictions were a proof that the hypothesis was valuable and, at least to a great extent, true. (Whewell, 1847, 62–63)

Indeed, science often proceeds that way: Our best theories about the atmospheric system suggest that emissions of greenhouse gases such as CO_2 and Methane lead to global warming. That hypothesis has been confirmed by its successful predictions, such as shrinking arctic ice sheets, increasing global temperatures, its ability to backtrack temperature variations in the past, etc. The hypothetico-deductive concept of confirmation explicates the common idea of these and similar examples by stating that evidence confirms a hypothesis if we can derive it from the tested hypothesis, together with suitable background assumptions. H-D confirmation thus naturally aligns with the Popperian method for scientific inquiry which emphasizes the value of risky predictions, the need to test our scientific hypotheses as severely as possible, to derive precise predictions and to check them with reality.

An elementary account of H-D confirmation is defined as follows:

Hypothetico-Deductive (H-D) Confirmation *E H-D-confirms H* relative to background knowledge K if and only if

1. $H.K$ is consistent,
2. $H.K$ entails E ($H.K \models E$),
3. K alone does not entail E.

The explicit role of background knowledge can be used to circumvent the raven paradox along the lines that Hempel suggested. Neither $Ra.Ba$ nor $\neg Ba.\neg Ra$ confirms the hypothesis $H = \forall x : Rx \rightarrow Bx$, but Ba ("a is black") does so *relative to the background knowledge* Ra, and $\neg Ra$ ("a is no raven") does so *relative to the background knowledge* $\neg Ba$. This makes intuitive sense: Only if we know a to be a raven, the observation of his color is evidentially relevant; and only if a is known to be non-black, the observation that it is no raven supports the hypothesis that all ravens are black, in the sense of eliminating a potential falsifier.

While the H-D account of confirmation fares well with respect to the raven paradox, it has a major problem. *Irrelevant conjunctions* can be tacked to the hypothesis H while preserving the confirmation relation (Glymour, 1980a).

> **Tacking by Conjunction Problem:** If H is confirmed by a piece of evidence E (relative to any K), $H.X$ is confirmed by the same E for an arbitrary X that is consistent with H and K.

It is easy to see that this phenomenon is highly unsatisfactory: Assume that the wave nature of light is confirmed by Young's double slit experiment. According to the H-D account of confirmation, this implies that the following hypothesis is confirmed: 'Light is an electromagnetic wave and the star Sirius is a giant bulb.' This sounds completely absurd.

To see that H-D confirmation suffers from the tacking problem, let us just check the three conditions for H-D confirmation: Assume that some hypothesis X is irrelevant to E, and that $H.X.K$ is consistent. Let us also assume $H.K \models E$ and that K alone does not entail E. Then, E confirms not only H, but also $H.X$ (because $H.K \models E$ implies $H.K.X \models E$).

Thus, tacking an arbitrary irrelevant conjunct to a confirmed hypothesis preserves the confirmation relation. This is very unsatisfactory. More generally, H-D confirmation needs an answer to why a piece of evidence does not confirm every theory that implies it. Solving this problem is perhaps not impossible (Schurz, 1991; Gemes, 1993; Sprenger, 2013), but comes at the expense of major technical complications that compromise the simplicity and intuitive appeal of the hypothetico-deductive approach of confirmation.

In our discussion, several problems of qualitative confirmation have surfaced. First, qualitative confirmation is grounded on deductive relations between theory and evidence. These are quite an exception in modern, statistics-based science which standardly deals with messy bodies of evidence. Second, we saw that few adequacy conditions have withstood the test of time, making times hard for developing a qualitiative *logic* of induction. Third, no qualitative account measures **degree of confirmation** and tells strongly from weakly confirmed hypotheses, although this is essential for a great deal of scientific reasoning. Therefore we now turn to quantitative explications of confirmation.

3 Probability as Degree of Confirmation

The use of probability as a tool for describing degree of confirmation can be motivated in various ways. Here are some major reasons.

First, probability is, as quipped by Cicero, "the guide to life". Judgments of probability motivate our actions: e.g., the train I want to catch will probably be on time, so I have to run to catch it. Probability is used for expressing forecasts about events that affect our lives in manifold ways, from tomorrow's weather to global climate, from economic developments to the probability of a new Middle East crisis. This paradigm was elaborated by philosophers and scientists such as Ramsey (1926), De Finetti (1937) and Jeffrey (1965).

Second, probability is the preferred tool for uncertain reasoning in science. Probability distributions are used for characterizing the value of a particular physical quantity or for describing measurement error. Theories are assessed on the basis of probabilistic hypothesis tests. By phrasing confirmation in terms of probability, we hope to connect philosophical analysis of inductive inference to scientific practice and integrate the goals of normative and descriptive adequacy (Howson and Urbach, 2006).

Third, statistics, the science of analysing and interpreting data, is couched in probability theory. Statisticians have proved powerful mathematical results on the foundations of probability and inductive learning. Analyses of confirmation may benefit from them, and have done so in the past (e.g., Good, 2009). Consider, for example, the famous De Finetti (1974) representation theorem for subjective probability or the convergence results for prior probability distributions by Gaifman and Snir (1982).

Fourth and last, increasing the probability of a conclusion seems to be the hallmark of a sound inductive inference, as already noted by Hume. Probability theory, and the Bayesian framework in particular, are especially well suited for capturing this intuition. The basic idea is to explicate degree of confirmation in terms of degrees of belief, which satisfy the axioms of probability. Degrees of belief are changed by Conditionalization (if E is learned, $p_{\text{new}}(H) = p(H|E)$), and the posterior probability $p(H|E)$ stands as the basis of inference and decision-making. This quantity can be calculated via Bayes' Theorem:

$$p(H|E) = p(H)\frac{p(E|H)}{p(E)}$$

The chapter on Probabilism provides more detail on the foundations of Bayesianism.

We now assume that degree of confirmation only depends on the joint probability distribution of the hypothesis H, the evidence E and the background assumptions K. More precisely, we assume that E, H and K are among the closed sentences $\mathfrak{L}$ of a language $\mathcal{L}$ that describes our domain of interest. A Bayesian theory of confirmation can be explicated by a function $\mathfrak{L}^3 \times \mathfrak{P} \to \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the algebra generated by $\mathfrak{L}$. This function assigns a real-valued degree of confirmation to any triple of sentences together with a probability (degree of belief) function. For the sake of simplicity, we will omit explicit reference to background knowledge since

most accounts incorporate it by using the probability function $p(\cdot|K)$ instead of $p(\cdot)$.

A classical method for explicating degree of confirmation is to specify **adequacy conditions** on the concept and to derive a **representation theorem** for a confirmation measure. This means that one characterizes the set of measures (and possibly the unique measure) that satisfies these constraints. This approach allows for a sharp demarcation and mathematically rigorous characterization of the explicandum, and at the same time for critical discussion of the explicatum, by means of defending and criticizing the properties which are encapsulated in the adequacy conditions.

The first constraint is mainly of formal nature and serves as a tool for making further constraints more precise and facilitating proofs (Crupi, 2013):

Formality For any sentences $H, E \in \mathfrak{L}$ with probability measure $p(\cdot)$, $c(H, E)$ is a measurable function from the joint probability distribution of H and E to a real number $c(H, E) \in \mathbb{R}$. In particular, there exists a function $f : [0, 1]^3 \to \mathbb{R}$ such that $c(H, E) = f(p(H \wedge E), p(H), p(E))$.

Since the three probabilities $p(H \wedge E)$, $p(H)$, $p(E)$ suffice to determine the joint probability distribution of H and E, we can express $c(H, E)$ as a function of these three arguments.

Another cornerstone for Bayesian explications of confirmation is the following principle:

Final Probability Incrementality For any sentences H, E, and $E' \in \mathfrak{L}$ with probability measure $p(\cdot)$,

$$\begin{array}{llll} c(H,E) > c(H,E') & \text{if and only if} & p(H|E) > p(H|E'), & \text{and} \\ c(H,E) < c(H,E') & \text{if and only if} & p(H|E) < p(H|E'). & \end{array}$$

According to this principle, E confirms H more than E' does if it raises the probability of H to a higher level. Given the basic intuition that degree of confirmation should co-vary with boost in degree of belief, satisfactory Bayesian explications of degree of confirmation should arguably satisfy this condition.

There are now two main roads for adding more conditions, which will ultimately lead us to two different explications of confirmation: as **firmness** and as **increase in firmness** (or evidential support). They are also called the **absolute** and the **incremental concept of confirmation**.

Consider the following condition:

Local Equivalence For any sentences H, H', and $E \in \mathfrak{L}$ with probability measure $p(\cdot)$, if H and H' are logically equivalent given E (i.e., $E.H \models H'$, $E.H' \models H$), then $c(H, E) = c(H', E)$.

The plausible idea behind Local Equivalence is that E confirms the hypotheses H and H' to an equal degree if they are logically equivalent conditional on E. If we buy into this intuition, Local Equivalence allows for a powerful (yet unpublished) representation theorem by Michael Schippers (see Crupi, 2013):

Theorem 1 Formality, Final Probability Incrementality and Local Equivalence hold if and only if there is a non-decreasing function $g : [0, 1] \to \mathbb{R}$ such that for any $H, E \in \mathfrak{L}$ and any $p(\cdot)$, $c(H, E) = g(p(H|E))$.

On this account, scientific hypotheses count as well-confirmed whenever they are sufficiently probable, that is, when $p(H|E)$ exceeds a certain (possibly context-relative) threshold. Hence, all confirmation measures that satisfy the three above constraints are *ordinally equivalent*, that is, they can be mapped on each other by means of a non-decreasing function. In particular, their confirmation rankings agree: if there are two functions g and g' that satisfy Theorem 1, with associated confirmation measures c and c', then $c(H, E) \geq c(H', E')$ if and only if $c'(H, E) \geq c'(H', E')$. Since confirmation as firmness is a monotonically increasing function of $p(H|E)$, it is natural to set up the qualitative criterion that E confirms H (in the absolute sense) if and only if $p(H|E) \geq t$ for some $t \in [0, 1]$.

A nice consequence of the view of confirmation as firmness is that some longstanding problems of confirmation theory, such as the paradox of irrelevant conjunctions, dissolve. Remember that on the H-D account of confirmation, it was hard to avoid the conclusion that if E confirmed H, then it also confirmed $H \wedge H'$ for an arbitrary H'. On the view of confirmation as firmness, we automatically obtain $c(H \wedge H', E) \leq c(H, E)$. These quantities are non-decreasing functions of $p(H \wedge H'|E)$ and $p(H|E)$, respectively, and they differ the more the less plausible H' is, and the less it coheres with H. Confirmation as firmness gives the intuitively correct response to the tacking by conjunction paradox.

It should also be noted that the idea of confirmation as firmness corresponds to Carnap's concept of probability$_1$ or "degree of confirmation" in his inductive logic. Carnap (1950) defines the degree of confirmation of a theory H relative to total evidence E as its probability conditional on E:

$$c(H, E) := p(H|E) = \frac{m(H \wedge E)}{m(E)}$$

where this probability is in turn defined by the measure m that descriptions of the (logical) universe receive. By the choice of the measure m and a learning parameter λ, Carnap (1952) characterizes an entire **continuum of inductive methods** from which three prominent special cases can be derived. First, *inductive skepticism*: the degree of confirmation of a hypothesis is not changed by incoming evidence. Second, the rule of *direct inference*: the degree of confirmation of the hypothesis equals the proportions of observations in the sample for which it is true. Third, the *rule of succession* (de Laplace, 1814), a prediction principle which corresponds to Bayesian inference with a uniform prior distribution. Carnap thus ends up with various inductive logics that characterize different attitudes toward ampliative inference.

Carnap's characterization of degree of confirmation does not always agree with the use of that concept in scientific reasoning. Above all, a confirmatory piece of evidence often provides a good *argument* for a theory, even if the latter is unlikely. For instance, in the first years after Einstein invented the General Theory of Relativity (GTR), many scientists did not have a particularly high degree of belief in GTR because of its counterintuitive nature. However, it was agreed upon that GTR was well-confirmed by its predictive and explanatory successes, such as the bending of starlight by the sun and the explanation of the Mercury perihelion shift (Earman, 1992). The account of confirmation as firmness fails to capture this intuition. The same holds for experiments in present-day science whose confirmatory strength is not evaluated on the basis of the posterior probability of the tested hypothesis H, but by whether the results provide significant evidence in favor of H, that is, whether they are more expected under H than under $\neg H$.

This last point brings us to a particularly unintuitive consequence of confirmation as firmness: E could confirm H even if it *lowers* the probability of H, as long as $p(H|E)$ is still large enough. But nobody would call an experiment where the results E are negatively statistically relevant to H a confirmation of H. This brings us to the following natural definition:

Confirmation as increase in firmness For any sentences H, $E \in \mathfrak{L}$ with probability measure $p(\cdot)$,

1. Evidence E **confirms/supports** hypothesis H (in the incremental sense) if and only if $p(H|E) > p(H)$.
2. Evidence E **disconfirms/undermines** hypothesis H if and only if $p(H|E) < p(H)$.
3. Evidence E is **neutral** with respect to H if and only if $p(H|E) = p(H)$.

	W_1	W_2
Black ravens	100	1,000
Non-black ravens	0	1
Other birds	1,000,000	1,000,000

Table 1: I.J. Good's (1967) counterexample to the paradox of the ravens.

In other words, E confirms H if and only E raises our degree of belief in H. Such explications of confirmation are also called **statistical relevance** accounts of confirmation because the neutral point is determined by the statistical independence of H and E. The analysis of confirmation as increase in firmness is the core business of **Bayesian Confirmation Theory**, where the relevant probabilities are interpreted as subjective degrees of belief. This approach receives empirical support from findings by Tentori et al. (2007): ordinary people use the concept of confirmation in a way that can be dissociated from posterior probability and that is strongly correlated with measures of confirmation as increase in firmness.

Confirmation as increase in firmness has interesting relations to qualitative accounts of confirmation and the paradoxes we have encountered. For instance, H-D confirmation now emerges as a special case: if H entails E, then $p(E|H) = 1$ and by Bayes' Theorem, $p(H|E) > p(H)$ (unless $p(E)$ was equal to one in the first place). We can also spot what is wrong with the idea of instance confirmation. Remember Nicod's (and Hempel's) original idea, namely that universal generalizations such as $H = \forall x : Rx \rightarrow Bx$ are confirmed by their instances. This is certainly true relative to *some* background knowledge. However, it is not true under *all* circumstances. I.J. Good (1967) constructed a simple counterexample in a note for the *British Journal for the Philosophy of Science*: There are only two possible worlds, W_1 and W_2, whose properties are described by Table 1.

Thus, H is true whenever W_1 is the case, and false whenever W_2 is the case. Conditional on these peculiar background assumptions, the observation of a black raven is evidence that W_2 is the case and therefore evidence that not all ravens are black:

$$P(Ra.Ba|W_1) = \frac{100}{1,000,100} \quad < \quad \frac{1,000}{1,001,001} = P(Ra.Ba|W_2).$$

By an application of Bayes' Theorem, we infer $P(W_1|Ra.Ba) < P(W_1)$, and given $W_1 \equiv H$, this amounts to a counterexample to Nicod's Condition (NC). Universal conditionals are not always confirmed by their positive instances. We see how confirmation as increase in firmness elucidates our pre-theoretic intuitions regarding the theory-evidence relation, and that the relevant background assumptions make a huge

difference as to when a hypothesis is confirmed.

Confirmation as increase in firmness also allows for a solution of the *comparative* paradox of the ravens. That is, we can show that relative to weak and plausible background assumptions, $p(H|Ra.Ba) < p(H|\neg Ra.\neg Ba)$ (Fitelson and Hawthorne, 2011, Theorem 2). By Final Probability Incrementality, this implies that $Ra.Ba$ confirms H more than $\neg Ra.\neg Ba$ does. This shows, ultimately, why we consider a black raven to be more important evidence for the raven hypothesis than a white shoe.

Looking back to qualitative accounts once more, we see that Hempel's original adequacy criteria are mirrored in the logical properties of confirmation as firmness and increase in firmness. According to the view of confirmation as firmness, every consequence H' of a confirmed hypothesis H is confirmed, too (because $p(H') \geq p(H)$). This conforms to Hempel's Special Consequence Condition. The view of confirmation as increase in firmness relinquishes this condition, however, and obtains a number of attractive results in return.

4 Degree of Confirmation: Monism or Pluralism?

So far, we have not yet answered the question of how degree of confirmation (or evidential support) should be quantified. For scientists who want to report the results of their experience and quantify the strength of the observed evidence, this is certainly the most interesting question. It is also crucial for giving a Bayesian answer to the Duhem-Quine problem (Duhem, 1914). If an experiment fails and we ask ourselves which hypothesis to reject, the degree of (dis)confirmation of the involved hypotheses can be used to evaluate their standing. Unlike purely qualitative accounts of confirmation, a measure of degree of confirmation can indicate which hypothesis we should discard. For this reason, the search for a proper confirmation measure is more than a technical exercise: it is of vital importance for distributing praise and blame between different hypotheses that are involved in an experiment. The question, however, is which measure should be used. This is the questions separating **monists** and **pluralists in confirmation theory**: monists believe that there is a single adequate or superior measure—a view that can be supported by theoretical reasons (Milne, 1996; Crupi et al., 2007) and empirical research, e.g., coherence with folk confirmation judgments (Tentori et al., 2007). Pluralists think that such arguments do not specify a single adequate measure and that there are several valuable and irreducible confirmation measures (e.g., Fitelson, 1999, 2001; Eells and Fitelson, 2000).

Table 2 provides a rough survey of the measures that are frequently discussed in the literature. We have normalized them such that for each measure $c(H, E)$, con-

Difference Measure	$d(H,E) = p(H\|E) - p(H)$
Log-Ratio Measure	$r(H,E) = \log \frac{p(H\|E)}{p(H)}$
Log-Likelihood Measure	$l(H,E) = \log \frac{p(E\|H)}{p(E\|\neg H)}$
Kemeny-Oppenheim Measure	$k(H,E) = \frac{p(E\|H)-p(E\|\neg H)}{p(E\|H)+p(E\|\neg H)}$
Rips Measure	$r'(H,E) = \frac{p(H\|E)-p(H)}{1-p(H)}$
Crupi-Tentori Measure	$z(H,E) = \begin{cases} \frac{p(H\|E)-p(H)}{1-p(H)} & \text{if } p(H\|E) \geq p(H) \\ \frac{p(H\|E)-p(H)}{p(H)} & \text{if } p(H\|E) < p(H) \end{cases}$
Christensen-Joyce Measure	$s(H,E) = p(H\|E) - p(H\|\neg E)$
Carnap's Relevance Measure	$c'(H,E) = p(H \wedge E) - p(H)p(E)$

Table 2: A list of popular measures of evidential support.

firmation amounts to $c(H,E) > 0$, neutrality to $c(H,E) = 0$ and disconfirmation to $c(H,E) < 0$. This allows for a better comparison of the measures and their properties.

Evidently, these measures all have quite distinct properties. We shall now transfer the methodology from our analysis of confirmation as firmness, and characterize them in terms of representation results. As before, Formality and Final Probability Incrementality will serve as minimal reasonable constraints on any measure of evidential support. Notably, two measures in the list, namely c' and s, are incompatible with Final Probability Incrementality, and objections based on allegedly vicious symmetries have been raised against c' and r (Fitelson, 2001; Eells and Fitelson, 2002).

Here are further constraints on measures of evidential support that exploit the increase of firmness intuition in different ways:

Disjunction of Alternatives If H and H' are mutually exclusive, then

$$c(H,E) > c(H \vee H', E') \qquad \text{if and only if} \qquad p(H'|E) > p(H'),$$

with corresponding conditions for $c(H,E) = c(H \vee H', E')$ and $c(H,E) < c(H \vee H', E')$.

That is, E confirms $H \vee H'$ more than H if and only if E is statistically relevant to H'. The idea behind this condition is that the sum $(H \vee H')$ is confirmed to a greater degree than each of the parts (H, H') when each part is individually confirmed by E.

Law of Likelihood

$$c(H,E) > c(H',E) \qquad \text{if and only if} \qquad p(E|H) > p(E|H'),$$

with corresponding conditions for $c(H, E) = c(H', E')$ and $c(H, E) < c(H', E')$.

This condition has a long history of discussion in philosophy and statistics (e.g., Hacking, 1965; Edwards, 1972). The idea is that E favors H over H' if and only the likelihood of H on E is greater than the likelihood of H' on E. In other words, E is more expected under H than under H'. Law of Likelihood also stands at the basis of the *likelihoodist theory of confirmation*, which analyzes confirmation as a comparative relation between two competing hypotheses (Royall, 1997; Sober, 2008). Likelihoodists eschew judgments on how much E confirms H without reference to specific alternatives.

Modularity If $p(E|H \wedge E') = p(E|H)$ and $p(E|\neg H \wedge E') = p(E|\neg H)$, then $c(H, E) = c_{|E'}(H, E)$ where $c_{|E'}$ denotes confirmation relative to the probability distribution conditional on E'.

This constraint screens off irrelevant evidence. If E' does not affect the likelihoods of H and $\neg H$ on E, then conditioning on E'—now supposedly irrelevant evidence—does not alter the degree of confirmation.

Contraposition/Commutativity If E confirms H, then $c(H, E) = c(\neg E, \neg H)$; and if E disconfirms H, then $c(H, E) = c(E, H)$.

These constraints are motivated by the analogy of confirmation to partial deductive entailment. If $H \vdash E$, then also $\neg E \vdash \neg H$, and if E refutes H, then H also refutes E. If confirmation is thought of as a generalization of deductive entailment to uncertain inference, then these conditions are very natural and reasonable (Tentori et al., 2007).

Combined with Formality and Final Probability Incrementality, each of these four principles singles out a specific measure of confirmation, up to ordinal equivalence (Heckerman, 1988; Crupi et al., 2013; Crupi, 2013):

Theorem 2 (Representation Results for Confirmation Measures) .

1. If Formality, Final Probability Incrementality and Disjunction of Alternatives hold, then there is a non-decreasing function g such that $c(H, E) = g(d(H, E))$.

2. If Formality, Final Probability Incrementality and Law of Likelihood hold, then there is a non-decreasing function g such that $c(H, E) = g(r(H, E))$.

3. If Formality, Final Probability Incrementality and Modularity hold, then there are non-decreasing functions g and g' such that $c(H, E) = g(l(H, E))$ and $c(H, E) = g'(k(H, E))$. Note that k and l are ordinally equivalent.

4. If Formality, Final Probability Incrementality and Commutativity hold, then there is a non-decreasing function g such that $c(H, E) = g(z(H, E))$.

That is, many confirmation measures can be chacterized by means of a small set of adequacy conditions. It should also be noted that the **Bayes factor**, a popular measure of evidence in Bayesian statistics (Kass and Raftery, 1995), falls under the scope of the theorem since it is ordinally equivalent to the log-likelihood measure l and the Kemeny and Oppenheim (1952) measure k. This is also evident from its mathematical form

$$\mathrm{BF}(H_0, H_1, E) := \frac{p(H_0|E)}{p(H_1|E)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{p(E|H_0)}{p(E|H_1)}$$

for mutually exclusive hypotheses H_0 and H_1 (for which H and $\neg H$ may be substituted).

To show that the difference between these measures has substantial philosophical ramifications, let us go back to the problem of irrelevant conjunctions. If we analyze this problem in terms of the ratio measure r, then we obtain, assuming $H \vdash E$, that for an "irrelevant" conjunct H',

$$\begin{aligned} r(H \wedge H', E) &= p(H \wedge H'|E)/p(H \wedge H') = p(E|H \wedge H')/p(E) \\ &= 1/p(E) = p(E|H)/p(E) \\ &= r(H, E) \end{aligned}$$

such that the irrelevant conjunction is supported to the same degree as the original hypothesis. This consequence is certainly unacceptable as a judgment of evidential support since H' could literally be any hypothesis unrelated to the evidence, e.g., "the star Sirius is a giant light bulb". In addition, the result does not only hold for the special case of deductive entailment: it holds *whenever the likelihoods of H and $H \wedge H'$ on E are the same*, that is, $p(E|H \wedge H') = p(E|H)$.

The other measures fare better in this respect: whenever $p(E|H \wedge H') = p(E|H)$, all other measures in Theorem 2 reach the conclusion that $c(H \wedge H', E) < c(H, E)$ (Hawthorne and Fitelson, 2004). In this way, we can see how Bayesian Confirmation Theory improves on H-D confirmation and other qualitative accounts of confirmation: the paradox is acknowledged, but at the same time, it is demonstrated how it can be mitigated.

That said, it is difficult to form preferences over the remaining measures. Comparing the adequacy conditions might not lead to conclusive results, due to the divergent motivations which support them. Moreover, it has been shown that none of the remaining measures satisfies the following two conditions: (i) degree of confirmation is maximal if E implies H; (ii) the a priori informativity (cashed out in terms of predictive content and improbability) of a hypothesis contributes to degree of confirmation (Brössel, 2013, 389–390). This means that the idea of confirmation as a generalization

of partial entailment and as a reward for risky predictions cannot be reconciled with each other, posing a further dilemma for confirmation monism. One may therefore go for pluralism, and accept that there are different senses of degree of confirmation that correspond to different explications. For example, d strikes us as a natural explication of increase in subjective confidence, z generalizes deductive entailment, and l and k measure the discriminatory force of the evidence regarding H and $\neg H$.

Although Bayesian Confirmation Theory yields many interesting results and has sparked interests among experimental psychologists, too, one main criticism has been levelled again and again: that it misrepresents actual scientific reasoning. In the remaining sections, we present two major challenges for Bayesian Confirmation Theory fed by that feeling: the Problem of Old Evidence (Glymour, 1980b) and the rivalling frequentist approach to learning from experience (Mayo, 1996).

5 The Problem of Old Evidence

In this brief section, we shall expose one of the most troubling and persistent challenges for confirmation as increase in firmness: the **Problem of Old Evidence**. Consider a phenomenon E that is unexplained by the available scientific theories. At some point, a theory H is discovered that accounts for E. Then, E is "old evidence": at the time when H is developed, the scientist is already certain or close to certain that the phenomenon E is real. Nevertheless, E apparently confirms H—at least if H was invented on independent grounds. After all, it resolves a well-known and persistent observational anomaly.

A famous case of old evidence in science is the Mercury perihelion anomaly (Glymour, 1980b; Earman, 1992). For a long time, the shift of the Mercury perihelion could not be explained by Newtonian mechanics or any other reputable physical theory. Then, Einstein realized that his General Theory of Relativity (GTR) explained the perihelion shift. This discovery conferred a substantial degree of confirmation on GTR, much more than some pieces of novel evidence. Similar reasoning patterns apply in other scientific disciplines where new theories explain away well-known anomalies.

The reasoning of these scientists is hard to capture in the Bayesian account of confirmation as increase in firmness. E confirms H if and only if the posterior degree of belief in H, $p(H|E)$, exceeds the prior degree of belief in H, $p(H)$. When E is old evidence and already known to the scientist, the prior degree of belief in E is maximal: $p(E) = 1$. But with that assumption, it follows that the posterior probability of H cannot be greater than the prior probability: $p(H|E) = p(H) \cdot p(E|H) \leq p(H)$. Hence, E does not confirm H. The very idea of confirmation by old evidence, or equivalently,

confirmation by accounting for well-known observational anomalies, seems impossible to describe in the Bayesian belief kinematics. Some critics, like Clark Glymour, have gone so far to claim that Bayesian confirmation only describes *epiphenomena* of genuine confirmation because it misses the relevant structural relations between theory and evidence.

There are various solution proposals to the Problem of Old Evidence. One approach, adopted by Howson (1984), interprets the confirmation relation with respect to counterfactual degrees of belief, where E is subtracted from the agent's actual background knowledge. Another approach is to claim that confirmation by old evidence is not about learning the actual evidence, but about **learning a logical or explanatory relation between theory and evidence**. It seems intuitive that Einstein's confidence in GTR increased upon learning that it implied the perihelion shift of Mercury, and that this discovery was the real confirming event.

Indeed, confirmation theorists have set up Bayesian models where learning $H \vdash E$ increases the probability of H (e.g., Jeffrey, 1983) under certain assumptions. The question is, of course, whether these assumptions are sufficiently plausible and realistic. For critical discussion and further constructive proposals, see Earman (1992) and Sprenger (2015a).

6 Bayesianism and Frequentism

A major alternative to Bayesian Confirmation Theory is **frequentist inference**. Many of its principles have been developed by the geneticist and statistician R.A. Fisher (see Neyman and Pearson, 1933, for a more behavioral account). According to frequentism, inductive inference does not concern our degrees of belief. That concept is part of individual psychology and not suitable for quantifying scientific evidence. Instead of expressing degrees of belief, probability is interpreted as the limiting frequency of an event in a large number of trials. It enters inductive inference via the concept of a **sampling distribution**, that is, the probability distribution of an observable in a random sample.

The basic method of frequentist inference is hypothesis testing, and more precisely, **null hypothesis significance tests** (NHST). For Fisher, the purpose of statistical analysis consists in assessing the relation of a hypothesis to a body of observed data. The tested hypothesis usually stands for the absence of an interesting phenomenon, e.g., no causal relationship between two variables, no observable difference between two treatments, etc. Therefore it is often called the default or **null hypothesis** (or shortly, null). In remarkable agreement with Popper, Fisher states that the only purpose of an

experiment is to "give the facts a chance of disproving the null hypothesis" (Fisher, 1925, 16): the purpose of a test is to find evidence *against* the null. Conversely, failure to reject the null hypothesis does not imply positive evidence for the null (on this problem, see Popper, 1954; Sprenger, 2015b).

Unlike Popper (1959), Fisher aims at experimental and statistical *demonstrations* of a phenomenon. Thus, he needs a criterion for when an effect is real and not an experimental fabrication. He suggests that we should infer to such an effect when the observed data are too improbable under the null hypothesis:

> "either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true." (Fisher, 1956, 39)

This basic scheme of inference is called **Fisher's Disjunction** by (Hacking, 1965), and it stands at the heart of significance testing. It infers to the falsity of the null hypothesis as the best explanation of an unexpected result (for criticism, see Spielman, 1974; Royall, 1997).

Evidence against the null is measured by means of the the **p-value**. Here is an illustration. Suppose that we want to test whether the real-valued parameter θ, our quantity of interest, diverges "significantly" from $H_0 : \theta = \theta_0$. We collect i.i.d. data $x := (x_1, \ldots, x_N)$ whose distribution is Gaussian and centered around θ. Assume now that the population variance σ^2 is known, so $x_i \sim N(\theta, \sigma^2)$ for each x_i. Then, the discrepancy in the data x with respect to the postulated mean value θ_0 is measured by means of the statistic

$$z(x) := \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N \cdot \sigma^2}}$$

We may re-interpret this equation as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}}$$

Determining whether a result is significant or not depends on the p-value or **observed significance level**, that is, the "tail area" of the null under the observed data. This value depends on z and can be computed as

$$p_{\text{obs}} := p(|z(X)| \geq |z(x)|),$$

that is, as the probability of observing a more extreme discrepancy under the null than the one which is actually observed. Figure 1 displays an observed significance level $p = 0.04$ as the integral under the probability distribution function—a result that would typically count as substantial evidence against the null hypothesis ("$p < .05$").

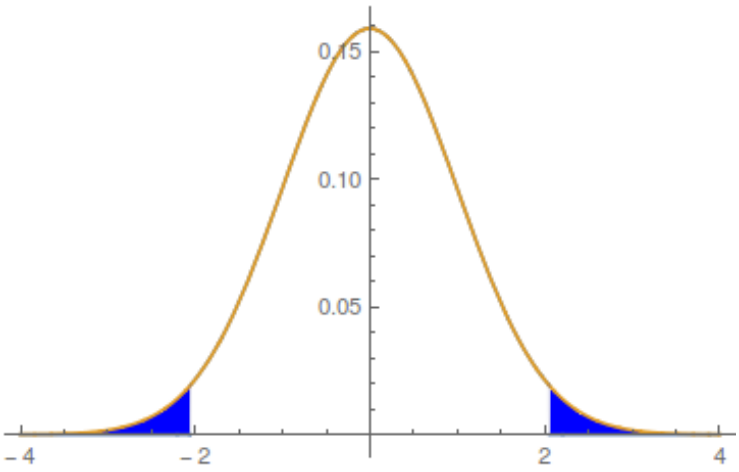

Figure 1: The probability density function of the null $H_0 : X \sim N(0,1)$, which is tested against the alternative $H_1 : X \sim N(\theta,1)$, $\theta \neq 0$. The shaded area illustrates the calculation of the p-value for an observed z-value of $z = \pm 2.054$ ($p = 0.04$).

For the frequentist practitioner, p-values are practical, replicable and objective measures of evidence against the null: they can be computed automatically once the statistical model is specified, and only depend on the sampling distribution of the data under H_0. Fisher interpreted them as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher, 1956, 43).

The virtues and vices of significance testing and p-values have been discussed at length in the literature (e.g., Cohen, 1994; Harlow et al., 1997), and it would go beyond the scope of this article to deliver a comprehensive critique. By now, it is consensus that inductive inference based on p-values leads to severe epistemic and practical problems. Several alternatives, such as **confidence intervals at a pre-defined level** α, have been promoted in recent years (Cumming and Finch, 2005; Cumming, 2013). They are interval estimators defined as follows: for each possible value θ' of the unknown parameter θ, we select the interval of data points x that will not lead to a statistical rejection of the null hypothesis $\theta = \theta'$ in a significance test at level α. Conversely, the confidence interval for θ, given an observation x, comprises all values of θ that are *consistent* with x in the sense of surviving a NHST at level α.

We conclude by highlighting the principal philosophical difference between Bayesian and frequentist inference. The following principle is typically accepted by Bayesian statisticians and confirmation theorists alike:

> **Likelihood Principle (LP):** Consider a statistical model $\mathcal{M}$ with a set of probability measures $p(\cdot|\theta)$ parametrized by a parameter of interest $\theta \in \Theta$.

> Assume we conduct an experiment $\mathcal{E}$ in $\mathcal{M}$. Then, all evidence about θ generated by $\mathcal{E}$ is contained in the *likelihood function* $p(x|\theta)$, where the observed data x are treated as a constant. (Birnbaum, 1962)

Indeed, in the simple case of only two hypotheses (H and $\neg H$), the posterior probabilities are only a function of $p(E|H)$ and $p(E|\neg H)$, given the prior probabilities. This is evident from writing the well-known Bayes' Theorem as

$$p(H|E) = \left(1 + \frac{p(\neg H)}{p(H)} \frac{p(E|\neg H)}{p(E|H)}\right)^{-1}$$

So Bayesians typically accept the LP, as is also evident from the use of Bayes factors as a measure of statistical evidence.

Frequentists reject the LP: his or her measures of evidence, such as p-values, are based on the probability of results that *could have happened, but did actually not happen*. The evidence depends on whether the actual data fit the null better or worse than most other possible data (see Figure 1). By contrast, Bayesian induction is "actualist": the only thing that matters for evaluating the evidence and making decisions is the predictive performance of the competing hypotheses on the actually observed evidence. Factors that determine the probability of possible, but unobserved outcomes, such as the experimental protocol, the intentions of the experimenter, the risk of early termination, etc., may have a role in experimental design, but they do not matter for measuring evidence *post hoc* (Edwards et al., 1963; Sprenger, 2009).

The Likelihood Principle is often seen as a strong argument for preferring Bayesian to frequentist inference (e.g., Berger and Wolpert, 1984). In practice, statistical data analysis still follows frequentist principles more often than not: mainly because in many applied problems, it is difficult to elicit subjective degrees of belief and to model prior probability distributions.

7 Conclusion

This chapter has given an overview of the problem of induction and the responses that philosophers of science have developed over time. These days, the focus is not so much on providing an answer to Hume's challenge: it is well-acknowledged that no purely epistemic, non-circular justification of induction can be given. Instead, focus has shifted to characterizing valid inductive inferences, carefully balancing attractive theoretical principles with judgments and intuitions in concrete cases. That this is not always easy has been demonstrated by challenges such as the paradox of the ravens, the problem of irrelevant conjunctions and Goodman's new riddle of induction.

In the context of this project, degree of confirmation becomes especially important: it indicates to what extent an inductive inference is justified. Explications of confirmation can be distinguished into two groups: qualitative and quantitative ones. The first serve well to illustrate the "grammar" of the concept, but they have limited applicability.

In Section 4 and 5, we have motivated why probability is an adequate tool for explicating degree of confirmation and investigated probabilistic (Bayesian) confirmation measures. We have distinguished two senses of confirmation—confirmation as firmness and confirmation as increase in firmness—and investigated various confirmation measures. That said, there are also alternative accounts of inductive reasoning, some of which are non-probabilistic, such as Objective Bayesianism (Williamson, 2010), ranking functions (Spohn, 1990), evidential probability (Kyburg, 1961) and the Dempster-Shafer theory of evidence (Shafer, 1976). See also Haenni et al. (2011).

Finally, we have provided a short glimpse of the methodological debate between Bayesians and frequentists in statistical inference. Confirmation theory will have to engage more and more with debates in statistical methodology if it does not want to lose contact to inductive inference in science—which was Bacon's target in the first place.

Suggested Readings

For qualiatitive confirmation theory, the classical texts are Hempel (1945a,b). For an overview of various logics of inductive inference with scientific applications, see Haenni et al. (2011). A classical introduction to Bayesian reasoning, with comparison to frequentism, is given by Howson and Urbach (2006). Earman (1992) and Crupi (2013) offer comprehensive reviews of Bayesian confirmation theory, and Good (2009) is an exciting collection of essays in induction, probability and statistical inference.

Acknowledgements

I would like to thank Matteo Colombo, Vincenzo Crupi, Raoul Gervais, Paul Humphreys, and Michael Schippers for their valuable feedback on this article. Research on this article was supported through the Vidi project "Making Scientific Inferences More Objective" (grant no. 276-20-023) by the Netherlands Organisation for Scientific Research (NWO).

References

Bacon, F. (1620). *Novum Organum; Or, True Suggestions for the Interpretation of Nature*. William Pickering, London.

Berger, J. and Wolpert, R. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward/CA.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, 57(298):269–306.

Broad, C. D. (1952). *Ethics and the History of Philosophy*. Routledge, London.

Brössel, P. (2013). The problem of measure sensitivity redux. *Philosophy of Science*, 80(3):378–397.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Carnap, R. (1952). *Continuum of Inductive Methods*. University of Chicago Press, Chicago.

Cohen, J. (1994). The Earth is Round (p<.05). *Psychological Review*, 49:997–1001.

Crupi, V. (2013). Confirmation. *The Stanford Encyclopedia of Philosophy*.

Crupi, V., Chater, N., and Tentori, K. (2013). New Axioms for Probability and Likelihood Ratio Measures. *British Journal for the Philosophy of Science*, 64(1):189–204.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues*. *Philosophy of Science*, 74:229–252.

Cumming, G. (2013). The New Statistics: Why and How . *Psychological Science*.

Cumming, G. and Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, 60(2):170–180.

De Finetti, B. (1937). La Prévision: ses Lois Logiques, ses Sources Subjectives. *Annales de l'institut Henri Poincaré*, 7:1–68.

De Finetti, B. (1974). *Theory of Probability*, volume 1. John Wiley & Sons, New York.

de Laplace, P. S. (1814). *A Philosophical Essay on Probabilities*. Dover, Mineola, NY.

Descartes, R. (1637). *Discours de la méthode*. Jan Maire, Leiden.

Duhem, P. (1914). *La Théorie Physique: Son Objet, Sa Structure*. Vrin, Paris.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, Mass.

Edwards, A. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Eells, E. and Fitelson, B. (2000). Measuring Confirmation and Evidence. *The Journal of Philosophy*, 97(12):663–672.

Eells, E. and Fitelson, B. (2002). Symmetries and Asymmetries in Evidential Support. *Philosophical Studies*, 107(2):129–142.

Fisher, R. (1956). *Statistical methods and scientific inference.* Hafner, New York.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh.

Fitelson, B. (1999). The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. In *Philosophy of Science*, volume 66, pages S362–S378.

Fitelson, B. (2001). *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin - Madison.

Fitelson, B. and Hawthorne, J. (2011). How Bayesian confirmation theory handles the paradox of the ravens. In Fetzer, J. H. and Eells, E., editors, *The Place of Probability in Science*, pages 247–275. Springer, New York.

Gaifman, H. and Snir, M. (1982). Probabilities Over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic*, 47(3):495–548.

Gemes, K. (1993). Hypothetico-Deductivism, Content and the Natural Axiomatisation of Theories. *Philosophy of Science*, 60:477–487.

Gemes, K. (1998). Hypothetico-deductivism: the current state of play; the criterion of empirical significance: endgame. *Erkenntnis*, 49(1):1–20.

Glymour, C. (1980a). Hypothetico-deductivism is hopeless. *Philosophy of Science*, 47(2):322–325.

Glymour, C. (1980b). *Theory and Evidence*. Princeton University Press, Princeton.

Goldman, A. I. (1986). *Epistemology and Cognition*. Harvard University Press, Cambridge, MA.

Good, I. (2009). *Good Thinking*. Dover, Mineola, NY.

Good, I. J. (1967). The white shoe is a red herring. *The British Journal for the Philosophy of Science*, 17(4):322.

Goodman, N. (1955). *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, MA.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge University Press, Cambridge.

Haenni, R., Romeijn, J.-W., Wheeler, G., and Williamson, J. (2011). *Probabilistic Logic and Probabilistic Networks*. Springer, Berlin.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (1997). *What if there were no significance tests?* Erlbaum, Mahwah/NJ.

Hawthorne, J. and Fitelson, B. (2004). Re-Solving Irrelevant Conjunction with Probabilistic Independence. *Philosophy of Science*, 71:505–514.

Heckerman, D. (1988). An Axiomatic Framework for Belief Updates. In J.F. Lemmer and L.N. Kanal, editor, *Uncertainty in Artificial Intelligence 2*, pages 11–22, Amsterdam. North-Holland.

Hempel, C. G. (1945a). Studies in the Logic of Confirmation {I}. *Mind*, 54(213):1–26.

Hempel, C. G. (1945b). Studies in the Logic of Confirmation {II}. *Mind*, 54(214):97–121.

Howson, C. (1984). Bayesianism and support by novel facts. *British Journal for the Philosophy of Science*, pages 245–251.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, IL, 3rd edition.

Hume, D. (1739). *A Treatise of Human Nature*. Clarendon Press, Oxford.

Hume, D. (1748). *Enquiry Concerning Human Understanding*. Clarendon Press, Oxford.

Jeffrey, R. C. (1965). *The Logic of Decision*. University of Chicago Press, Chicago, 2nd edition.

Jeffrey, R. C. (1983). Bayesianism with a Human Face. In Earman, J., editor, *Testing scientific theories*, pages 133–156. University of Minnesota Press, Minneapolis, minnesota edition.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90:773–795.

Kemeny, J. G. and Oppenheim, P. (1952). Degree of Factual Support. *Philosophy of Science*, 19:307–324.

Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief*. Wesleyan University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

Mayo, D. G. and Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *British Journal for the Philosophy of Science*, 57:323–357.

Milne, P. (1996). log[P(h/eb)/P(h/b)] is the One True Measure of Confirmation. *Philosophy of Science*, 63:21–26.

Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses.

Nicod, J. (1961). *Le problème logique de l'induction*. Presses Universitaires de France, Paris.

Norton, J. D. (2003). A Material Theory of Induction. *Philosophy of Science*, 70(4):647–670.

Popper, K. (1954). Degree of confirmation. *The British Journal for the Philosophy of Science*, 5:143–149.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London.

Popper, K. R. (1983). *Realism and the Aim of Science*. Rowman & Littlefield, Towota, NJ.

Ramsey, F. P. (1926). Truth and Probability. In Mellor, D. H., editor, *Philosophical Papers*, pages 52–94. Cambridge University Press, Cambridge.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Schurz, G. (1991). Relevant deduction. *Erkenntnis*, 35:391–437.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press, Cambridge.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science*, 41(3):211–226.

Spohn, W. (1990). A General Non-Probabilistic Theory of Inductive Reasoning. In Shachter, R. D., Levitt, T. S., Lemmer, J., and Kanal, L. N., editors, *Uncertainty in Artificial Intelligence 4*. Elsevier, Amsterdam.

Sprenger, J. (2009). Evidence and Experimental Design in Sequential Trials. *Philosophy of Science*, 76:637–649.

Sprenger, J. (2010). Hempel and the Paradoxes of Confirmation. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Handbook of the History of Logic*, volume 10, pages 235–263. North-Holland, Amsterdam.

Sprenger, J. (2011). Hypothetico-Deductive Confirmation. *Philosophy Compass*, 6(7):497–508.

Sprenger, J. (2013). A Synthesis of Hempelian and Hypothetico-Deductive Confirmation. *Erkenntnis*, 78:727–738.

Sprenger, J. (2015a). A Novel Solution of the Problem of Old Evidence. *Philosophy of Science*.

Sprenger, J. (2015b). Two Impossibility Results for Measures of Corroboration.

Tentori, K., Crupi, V., Bonini, N., and Osherson, D. (2007). Comparison of Confirmation Measures. *Cognition*, 103:107–119.

Vickers, J. (2010). The Problem of Induction. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2010 edition.

Whewell, W. (1847). *Philosophy of the Inductive Sciences, Founded Upon Their History*. Parker, London.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press, Oxford.